Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps

Randy C. Shoemaker, David Grant, Terry Olson, Wesley C. Warren, Rod Wing, Yeisoo Yu, HyeRan Kim, Perry Cregan, Bindu Joseph, Montona Futrell-Griggs, Will Nelson, Jon Davito, Jason Walker, John Wallis, Colin Kremitski, Debbie Scheer, Sandra W. Clifton, Tina Graves, Henry Nguyen, Xiaolei Wu, Mingcheng Luo, Jan Dvorak, Rex Nelson, Steven Cannon, Jeff Tomkins, Jeremy Schmutz, Gary Stacey, and Scott Jackson

Abstract: Whole-genome sequencing of the soybean (*Glycine max* (L.) Merr. 'Williams 82') has made it important to integrate its physical and genetic maps. To facilitate this integration of maps, we screened 3290 microsatellites (SSRs) identified from BAC end sequences of clones comprising the 'Williams 82' physical map. SSRs were screened against 3 mapping populations. We found the AAT and ACT motifs produced the greatest frequency of length polymorphisms, ranging from 17.2% to 32.3% and from 11.8% to 33.3%, respectively. Other useful motifs include the dinucleotide repeats AG, AT, and AG, with frequency of length polymorphisms ranging from 11.2% to 18.4% (AT), 12.4% to 20.6% (AG), and 11.3% to 16.4% (GT). Repeat lengths less than 16 bp were generally less useful than repeat lengths of 40–60 bp. Two hundred and sixty-five SSRs were genetically mapped in at least one population. Of the 265 mapped SSRs, 60 came from BAC singletons not yet placed into contigs of the physical map. One hundred and ten originated in BACs located in contigs for which no genetic map location was previously known. Ninety-five SSRs came from BACs within contigs for which one or more other BACs had already been mapped. For these fingerprinted contigs (FPC) a high percentage of the mapped markers showed inconsistent map locations. A strategy is introduced by which physical and genetic map inconsistencies can be resolved using the preliminary 4× assembly of the whole genome sequence of soybean.

Key words: SSR, molecular marker, genome sequence, assembly, physical map.

Résumé : Le séquençage complet du génome du soya (*Glycine max* (L.) Merr. 'Williams 82') a rendu importante l'intégration des cartes génétique et physique. Afin de faciliter cette intégration, les auteurs ont criblé 3290 microsatellites (SSR) trouvés au sein des séquences terminales de clones BAC qui composent la carte physique de 'Williams 82'. Les SSR ont été examinés chez trois populations de cartographie. Les motifs AAT et ACT ont produit les fréquences les plus élevées de polymorphisme; pour ces deux motifs, le polymorphisme variait respectivement entre 17,2 % et 32,3 % ainsi qu'entre 11,8 % et 33,3 %. Parmi les autres motifs utiles, notons les répétitions dinucléotidiques AG, AT et GT dont les proportions de locus polymorphes étaient les suivantes : 11,2 % – 18,4 % (AT), 12,4 % – 20,6 % (AG) et 11,3 % – 16,4 % (GT). Les répétitions totalisant moins de 16 pb étaient généralement

Received 10 October 2007. Accepted 18 January 2008. Published on the NRC Research Press Web site at genome.nrc.ca on 18 March 2008.

Corresponding Editor: P. Gustafson.

R.C. Shoemaker,[1] D. Grant, R. Nelson, and S. Cannon. USDA-ARS-CICGR Unit, Department of Agronomy, Ames, IA 50011-1010, USA.
T. Olson and B. Joseph. Department of Agronomy, Iowa State University, Ames, IA 50011, USA.
W.C. Warren, J. Davito, J. Walker, J. Wallis, C. Kremitski, D. Scheer, S.W. Clifton, and T. Graves. Washington University School of Medicine, Genome Sequencing Center, St. Louis, MO 63108, USA.
R. Wing, Y. Yu, H. Kim, and W. Nelson. Arizona Genomics Institute, Department of Plant Sciences, BIO5, University of Arizona, Tucson, AZ 85721, USA; Arizona Genomics Computational Laboratory, BIO5, University of Arizona, Tucson, AZ 85271, USA.
P. Cregan. USDA-ARS, Soybean Genomics and Improvement Laboratory, Beltsville, MD 20705, USA.
M. Futrell-Griggs and S. Jackson. Department of Agronomy, Purdue University, Lafayette, IN 47907, USA.
H. Nguyen, X. Wu, and G. Stacey. National Center for Soybean Biotechnology, Department of Plant Sciences, University of Missouri, Columbia, MO 65211, USA.
M. Luo and J. Dvorak. Department of Plant Sciences, University of California, Davis, CA 95616, USA.
J. Tomkins. Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29631, USA.
J. Schmutz. Joint Genome Institute – Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, Palo Alto, CA 94304, USA.

[1] Corresponding author (e-mail: rcsshoe@iastate.edu).

doi:10.1139/G08-010

moins utiles que celles totalisant entre 40 et 60 pb. Deux cent cinquante-six SSR ont été cartographiés sur au moins une population. Parmi les 265 SSR cartographiés, 60 provenaient de clones BAC uniques ne faisant pas encore partie de contigs composant la carte physique. Cent dix provenaient de clones BAC appartenant à des contigs dont la position sur la carte génétique n'était pas encore connue. Quatre-vingt-quinze SSR provenaient de clones BAC appartenant à des contigs dont au moins un BAC avait déjà été situé sur la carte génétique. Parmi ces contigs FPC, une forte proportion des marqueurs cartographiés souffrait de positions incohérentes. Une stratégie est employée pour résoudre les incohérences entre les cartes physique et génétique à l'aide de l'assemblage préliminaire 4× du génome du soya.

Mots-clés : SSR, marqueur moléculaire, séquence du génome, assemblage, carte physique.

[Traduit par la Rédaction]

Introduction

A physical map of the soybean (*Glycine max* (L.) Merr.) cultivar 'Williams 82', the genotype "standard" chosen by the soybean community for genomic studies (Stacey et al. 2004), was developed using a high information content fingerprint (HCIF) approach (Luo et al. 2003; Warren and The Soybean Mapping Consortium 2006). The initial map was developed from two independent BAC libraries and was assembled with FPC (Soderlund et al. 2000; Pampanwar et al. 2005; Warren and The Soybean Mapping Consortium 2006). This map was derived from 97 272 fingerprinted BAC clones, which were assembled into 1893 contigs and approximately 30 000 singletons; this map is available at SoyBase (http://www.soybase.org) and The Soybean Breeder's Toolbox at http://soybeanphysicalmap.org/.

For an FPC map to be useful in the assembly of a whole genome sequence it must be anchored and overlaid onto the genetic map (Jackson et al. 2006). A high-quality physical map is an essential tool for improving and assessing the validity of a whole genome sequence assembly (Warren et al. 2005, 2006). The association of the physical map with the genetic map also has immediate application for cloning genes. In addition, a genetically anchored physical map of a complex genome allows better understanding of genome structure and organization, which may be misinterpreted using a strict whole-genome sequencing strategy.

Microsatellites (SSRs) are tandem repeat sequences of various lengths that are widely distributed throughout the genomes of plants and animals (Burow and Blake 1998). They are commonly detected through gel electrophoresis of a polymerase chain reaction (PCR) amplification product that encompasses the repeat sequence. The high frequency of length polymorphisms among different genotypes has made SSRs useful as molecular markers in a large number of species of plants and animals.

Microsatellites were first described for soybean by Akkaya et al. (1992). They were used for genetic mapping in soybean (Morgante et al. 1994; Akkaya et al. 1995) and then for assessment of variation among *G. max* and *G. soja* accessions (Powell et al. 1996). Soybean SSRs have been used with limited success across other legume genera (Peakall et al. 1998). Their ability to detect single loci in a paleopolyploid genome such as soybean made them essential for the creation of an integrated genetic linkage map (Cregan et al. 1999; Song et al. 2004). SSRs have allowed us to align the molecular-marker linkage map to the cytogenetic map (Zou et al. 2003) and position 'Forrest' BAC clones to the genetic map (Shultz et al. 2007).

The ability of SSRs to detect single loci, even in a paleopolyploid genome such as soybean, would seem to imply that a single SSR should be associated with a single locus on a map. However, for the 'Forrest' physical map it was reported that many DNA markers seemed to "anchor" several distinct contigs (Shultz et al. 2006). The same situation is observed in the 'Williams 82' physical map (http://soybeanphysicalmap.org). Improvement of the quality of a physical map requires accurate genetic anchoring of FPC contigs and a means by which contradictions in FPC contig map locations can be resolved.

In this paper we report on the overlaying of the soybean 'Williams 82' physical map onto the genetic map through identification and mapping of SSRs from end sequences of BACs integrated into the physical map. We also propose a strategy for the reciprocal quality control of the soybean genetic and physical maps and the forthcoming whole genome sequence assembly.

Methods and materials

BAC library construction

For map construction, two *Glycine max* 'Williams 82' BAC libraries were obtained. The GM_WBa BAC library (Marek and Shoemaker 1997; previously referred to as Gm_ISb001) was constructed using the *Hin*dIII site in pBeloBAC11. The library consists of approximately 40 320 clones with an average insert size of 116 kb, representing approximately 4 haploid genome equivalents. The GM_WBb library was constructed using the *Bst*YI site of pCUGIBAC1 (C. Saski and J. Tomkins, Clemson University Genomics Institute, unpublished data). The library consists of 67 968 BAC clones with an average insert size of 138 kb representing approximately 9 haploid genome equivalents.

BAC fingerprinting

Glycine max 'Williams 82' GM_WBa (*Hin*dIII) and GM_WBb (*Bst*YI) BAC libraries were fingerprinted as described by Luo et al. (2003) with minor modifications (below). In each 384-well plate, four blocks of 96 wells containing 1.2 mL of 2× YT medium (Sambrook et al. 1989) were inoculated with cells with a 96-well replicator. Two pins were removed from the replicator for the insertion of control clones into the 96-well blocks. Two control BAC clones were inserted manually in wells E07 and H12 in each 96-well block. The cultures were grown for 24 h on an orbital shaker agitated at 400 r/min at 37 °C. BAC DNAs were isolated with the QIAGEN R.E.A.L. Prep 96 Plasmid

Kit (QIAGEN, Valencia, California). The following minor modifications of the fingerprinting method were made to accommodate the use of the ABI 3730*xl* (Applied Biosystems, Foster City, California) instead of the ABI 3100 for capillary electrophoresis. To reduce sample size, 0.5–1.2 μg instead of 1.0–2.0 μg of BAC DNA was simultaneously digested with 2.0 instead of 5.0 units of each of *Bam*HI, *Eco*RI, *Xba*I, *Xho*I, and *Hae*III (New England Biolabs, Beverly, Massachusetts) at 37 °C for 3 h. DNAs were labeled with 0.4 μL instead of 1.0 μL of the SNaPshot Multiplex Ready Reaction Mix (Applied Biosystems) at 65 °C for 1 h and precipitated with ethanol. DNAs were dissolved in 9.9 μL of Hi-Di formamide, and 0.1 μL instead of 0.2 μL of GeneScan LIZ500 size standard was added to each sample. Restriction fragments were sized with an ABI 3730*xl* using 36 cm capillaries and POP-7 polymer (Applied Biosystems). Fragment size calling was accomplished with GeneMapper software (Applied Biosystems) with the help of FPPipeliner (http://www.bioinforsoft.com/). The GeneMapper output data were edited with the GenoProfiler program (You et al. 2007). The control BAC clones inserted in each 96-well block were used to check for the correct orientation of each plate. Fingerprints of cross-contaminated samples were detected using a module in GenoProfiler (You et al. 2007) and eliminated from the data set.

Physical map construction

Initial assembly parameters of individual clone fingerprints were determined empirically. After an optimized initial assembly of the fingerprints at a Sulston score of e^{-80}, tolerance of 4, we completed a series of contig merges with the automerge function of FPC (Soderlund et al. 2000; Pampanwar et al. 2005). Singletons were added up until a Sulston score of e^{-66} and once again at a score of e^{-45}, restricting additions to clones with a minimum of 5 clone matches. Once clone order was established, potential joins between contigs were identified by querying the local database with clones at the extreme ends of each contig, at a reduced fingerprint overlap stringency. At each stage of automerge, contigs were surveyed manually and Q clone statistics were used to monitor contig quality. All contig merges were stopped at a final score of e^{-26}. Attempts to reduce the score further resulted in significant increases in Q clone content. A final attempt to incorporate additional clones into the map was accomplished by building the singleton pool of clones as a separate map and adding clone contigs of 10 or greater. This build was done at a score of e^{-26}.

Identification of microsatellites

Fifty-nine thousand two-hundred and eighty-six BAC end sequences (BES) and 121 681 BES with an average insert size of 692 bp were retrieved from NCBI for the GM_WBa and GM_WBb libraries, respectively. All BES were generated by the Arizona Genomics Institute (H. Kim, Y. Yu, and R. Wing, unpublished data). These BES were searched for microsatellite repeats. Microsatellites (di-, tri-, tetra-, and penta-nucleotide core types, and mixed) were identified using two protocols: Sputnik (http://espressosoftware.com/pages/sputnik.jsp) and an in-house stepping and matching algorithm. The latter program will accept up to 60 MB of FASTA sequence input. The user specifies a threshold repeat size and the program reports both "perfect" and "imperfect" repeats (i.e., those containing mismatched bases as insertions, deletions, or mutations relative to the core repeat unit). Imperfect SSRs were accepted only if the mismatch count was less than or equal to 10% of total bases and the copy number count was at least the minimum allowed by the threshold repeat size. Only SSRs 16 bp or larger were evaluated.

PCR primers were chosen using Primer3 (Rozen and Skaletsky 2000). We empirically determined the minimal product size and highest primer quality that allowed Primer3 to find primers for over 90% of all targets.

Screening for length polymorphisms and genetic mapping

Microsatellites were evaluated for their ability to detect length polymorphisms among pairs of soybean cultivars or accessions: A81-356022 and PI 468.916, PI 437.654 and BSR 101, and Minsoy and Noir I. These pairs are the parents in previously developed mapping populations. PI 468.916 is a *G. soja* accession, while the others are *G. max*.

Genetic mapping of length polymorphisms was conducted in two populations. BSR 101 × PI 437.654 is a recombinant inbred line (RIL) population consisting of 320 individuals selected in the $F_{6:7}$ generation (Lewers et al. 1999). Ninety-four lines were selected for marker placement in this study. The population A81-356022 × PI 468.916 is an established interspecific population in the $F_{2:4}$ generation. It was first reported by Shoemaker and Olson (1993) and was used in the construction of the soybean composite map (Cregan et al. 1999; Song et al. 2004).

The gel system was 6% acrylamide in 0.5× TBE buffer pre-run at 300 V for 80 min with ethidium bromide. Samples were electrophoresed at 275 V for 120 min following the protocol of Wang et al. (2003). Marker placement was accomplished using MAPMAKER 3.0b (Lincoln et al. 1993).

Quality control of the genetic and physical map associations

The preliminary 4× sequence assembly was created from 7 851 733 reads at the Joint Genome Institute – Stanford Human Genome Center, using the Arachne assembler (Batzoglou et al. 2002; Jaffe et al. 2003). The sequence space represented 1 130 579 689 bases. Assembly resulted in sequence contigs with an average size of 157 040 bases and a maximum length of 20 109 437 bases. The preliminary assembly was made available to us by JGI for the purpose of this analysis.

We used the preliminary 4× whole genome assembly (4× WGS) of the soybean genome generated by JGI to compare the positions of the microsatellites themselves to those of the BACs they were associated with in earlier contig anchoring efforts. For previously reported microsatellites, the BAC end sequences (BES) for each BAC that contained a microsatellite and the genomic sequence from which the microsatellite was derived were searched for sequence homology using BLAST with cut-off values of $10e^{-50}$ and $10e^{-4}$, respectively (Altschul et al. 1997) against the 4× WGS. For the microsatellites, we used both of the BES from the BAC from which the microsatellite was derived. We considered

Table 1. Usefulness of repeat class types in detecting length polymorphisms between parents of three soybean mapping populations.

		Mapping population parental genotypes		
Repeat class	Mean size (bp)	A81-356022 vs. PI 468.916	BSR 101 vs. PI 437.654	Minsoy vs. Noir I
$(AT)_n$	43	18.4 (282/1534)	14.9 (181/1213)	11.2 (136/1213)
$(GT)_n$	43	16.4 (49/298)	15.4 (37/240)	11.3 (27/240)
$(AG)_n$	28	22.7 (75/330)	20.6 (50/243)	12.4 (30/243)
$(AAC)_n$	23	15.3 (31/203)	14.0 (21/150)	11.3 (17/150)
$(AAG)_n$	21	6.1 (28/458)	18.2 (16/195)	5.6 (11/195)
$(AAT)_n$	38	32.3 (132/409)	27.0 (88/326)	17.2 (56/326)
$(ACC)_n$	19	8.0 (2/25)	7.1 (1/14)	0 (0/14)
$(ACG)_n$	18	0 (0/2)	0 (0/1)	0 (0/1)
$(ACT)_n$	21	11.76 (2/17)	33.3 (4/12)	16.7 (2/12)
$(AGC)_n$	18	9.1 (1/11)	0 (0/7)	0 (0/7)
$(AGG)_n$	21	14.6 (8/55)	7.7 (3/39)	0 (0/39)
$(ATC)_n$	18	12.4 (10/81)	11.1 (3/27)	14.8 (4/27)
$(CCG)_n$	19	40.0 (2/5)	33.3 (1/3)	0 (0/3)

Note: Values shown are the frequency (%) of length polymorphism. In parentheses are the number of SSRs producing a length polymorphism detectable using the described gel conditions (before the slash) and the number of SSRs tested against the parental genotype (after the slash).

the top BLAST match (at least 98% identity) to the genome sequence scaffolds to indicate correspondence between the BAC and the WGS.

Mapping of all BES from a single FPC contig onto a WGS scaffold was taken as evidence that the FPC contig was likely assembled correctly. Similarly, correspondence of the genomic sequence from which previously mapped microsatellites had been derived (marker sequence) to the same WGS contig was taken as evidence that the microsatellite was properly associated with the FPC contig. Lack of correspondence between the WGS contig identified using marker sequence and the WGS contig identified by the BES putatively identified by the marker suggested that the microsatellite was falsely associated to the FPC contig.

Using this approach we assayed BES and marker sequence from 22 FPC contigs that contained at least two previously mapped molecular markers, but one or more of them with a position inconsistent with the linkage map. We also assayed BES and marker sequence from another 12 FPC contigs chosen at random that contained similar conflicting positions but did not contain any SSRs mapped in this study.

Results

Physical map

Previous RFLP hybridization to high-density filters and PCR screening of multidimensional pools of BAC clones using genetically mapped SSR primers resulted in identification of 472 contigs with genetic map positions. One hundred and nine of these contigs contained markers from more than one linkage group (http://soybeanphysicalmap.org/). This was not surprising given the hybridization-based protocol and the duplicated nature of the soybean genome (Shoemaker et al. 1996; Schlueter et al. 2004). The original build of the map can be found at SoyBase (http://www.soybase.org) and The Soybean Breeder's Toolbox (http://soybeanphysicalmap.org/). Manual curation of this map by several groups continues.

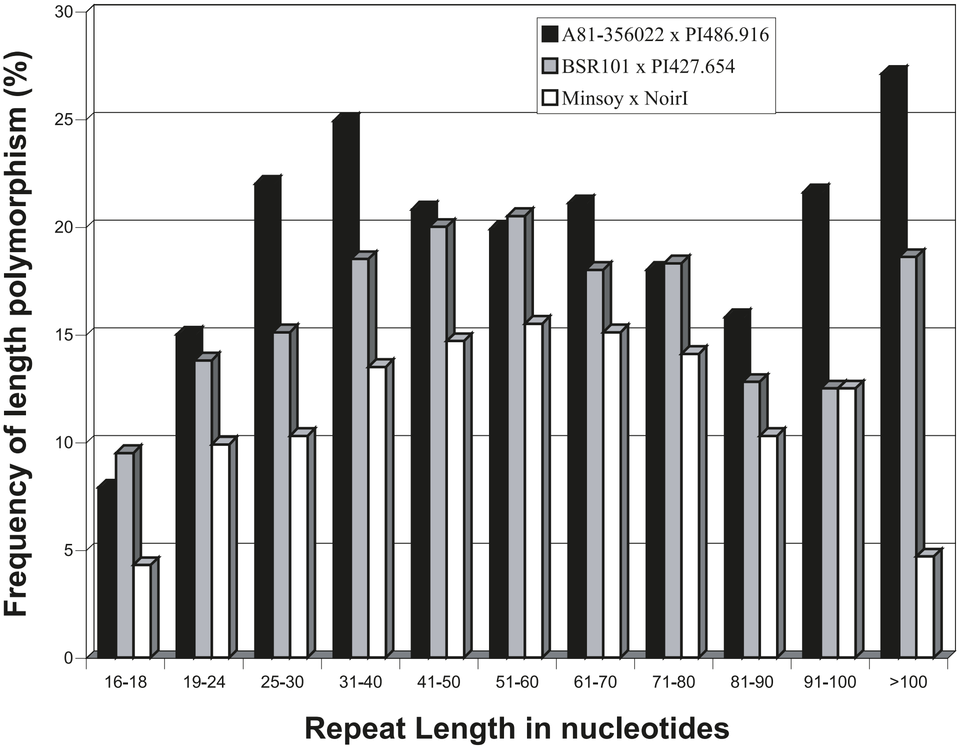
Effect of repeat class and repeat length on detection of length polymorphisms

Oligonucleotide primers were designed for a total of 3290 microsatellites. All microsatellites were screened against A81-356022 (*G. max*) and PI 468.916 (*G. soja*), parents of an interspecific mapping population. A portion of the microsatellites (2054) were screened against BSR 101 and PI 437.654 and against Minsoy and Noir I, parents of two additional intraspecific mapping populations.

There was a strong relationship between a microsatellite's repeat class and its usefulness as a genetic marker. The AAT class of repeats produced the greatest frequency of length polymorphisms, ranging from 17.2% in the Minsoy × Noir I cross to 32.3% in the interspecific cross (A81-356022 × PI 468.916) (Table 1). The ACT motif is also quite productive (Table 1), although it was not an abundant class. Other useful repeat classes include the dinucleotide repeats AT, AG, and GT, with frequency of length polymorphisms ranging from 11.2% to 18.4% (AT), 12.4% to 22.7% (AG), and 11.3% to 16.4% (GT). Surprisingly, some repeat classes, such as AGG, were very useful in the interspecific population but less informative in one or both of the intraspecific crosses, while others, such as AAG, were 3 times more informative in the BSR 101 × PI 437.654 population than in either of the other two populations (Table 1).

The repeat length (number of repeat unit repetitions) of a microsatellite was also strongly associated with its usefulness as a genetic marker. Repeat lengths less than 16 bp were not generally useful for detecting length polymorphisms. Dinucleotide repeats were generally longer and showed a wider range of length than trinucleotide repeats. Within the interspecific cross combination the frequency of length polymorphisms increased steadily as repeat length increased from 16 bp to 40 bp, where it peaked at 24%, then remained relatively constant through repeat lengths up to 70 bp, and then declined through lengths up to 90 bp. Frequency of length polymorphisms then increased to approximately 27% with repeat lengths greater than 100 bp (Fig. 1).

Fig. 1. The effect of microsatellite length on the frequency of detectable length polymorphisms between parents of three soybean mapping populations.



The frequency of length polymorphisms increased steadily in the intraspecific combinations as repeat length increased to 51–60 bp. The frequency of length polymorphisms between the Minsoy and Noir I parents then began to decline as repeat length increased. An increase in frequency of length polymorphisms was again seen with repeats greater than 100 bp in the BSR 101 × PI 437.654 cross combination (Fig. 1).

Genetic mapping of BES-derived SSRs

To improve the quality of the 'Williams 82' physical map, 265 SSRs discovered in BES were genetically mapped in at least one of two populations: A81-356022 × PI 468.916 and BSR 101 × PI 437.654. The distribution of the repeat classes of the mapped SSRs among the 20 soybean linkage groups is shown in Table 2. Di- and tri-nucleotide repeats were the most abundant and the most frequently mapped. Mononucleotide repeats were not assayed. Detailed information on SSR primer sequences, repeat motifs, associated linkage groups, and allele sizes can be found in the United States Department of Agriculture's soybean genetic database at http://soybase.org/RCS%20paper%20web%20table%201.html.

Of the 265 mapped SSRs, 60 were derived from BAC singletons, clones not yet placed into FPC contigs of the physical map. Eight of these 60 SSRs came from BACs already associated with a genetically mapped marker and the SSR confirmed the map position of the BAC, while the remaining 52 came from previously unmapped BACs. One hundred and ten of the 265 SSRs originated from BACs thought to be assigned to 90 distinct FPC contigs for which no genetic map location was previously known. This resulted in some contigs containing more than one BAC from which SSRs were developed and mapped. The last 95 SSRs came from BACs assigned to FPC contigs with at least one BAC associated with a mapped marker. Fifteen of the 67 contigs in this class contained SSR markers anchoring the contig to the linkage map. Twenty-four contigs associated with a linkage group contained SSR markers that mapped to different locations in the genome. This incongruity suggested that the assignments of some BACs warranted revision. Map positions of new SSRs did not match those of other previously mapped markers in an additional 27 contigs.

Quality control of the genetic and physical map associations

The association of all BAC end sequences from one FPC contig to one whole genome sequence contig was taken as evidence that the FPC BAC contig was likely assembled correctly. Correspondence of the genomic sequence from which a previously mapped microsatellite had been derived to the same WGS contig was taken as evidence that the microsatellite was properly associated with the FPC contig. Lack of correspondence between the WGS contig identified using the sequence from which the microsatellite had been derived and the WGS contig identified by the BES putatively identified by the marker indicated that the microsatellite was falsely associated to the FPC contig.

Using this approach we assayed BES and marker sequence from 22 physical contigs that contained 36 SSRs mapped in this study. These FPC contigs were chosen be-

Table 2. Distribution of repeat classes of mapped microsatellites on soybean linkage groups.

	Linkage group																			
Repeat unit size	Al	A2	B1	B2	C1	C2	D1a	D1b	D2	E	F	G	H	I	J	K	L	M	N	O
Di	8	2	7	11	12	7	11	8	7	7	7	5	12	11	2	6	8	8	15	7
Tri	8	5	3	2	3	4	8	1	5	7	3	2	2	3	5	5	5	6	6	2
Tetra				1		1			2					3		2				
Penta		1		3					1						1			1		1
Mixed	1			1																

Note: Repeat unit sizes range from 2 to 5 nucleotides. "Mixed" refers to a subset of instances wherein multiple repeats consisting of more than one core type are flanked by a single pair of primers.

Table 3. Examples of 'Williams 82' physical and genetic map quality control.

FPC contig No.	Marker	LG	Marker WGS[a]	BES-F WGS[b]	BES-R WGS[b]	BES-F WGS	BES-R WGS	BES-F WGS	BES-R WGS	BES-F WGS	BES-R WGS
3955	SSR1766	E	111[c]	111	111						
	Satt542	D1b	247	111	111						
	Satt268	E	111	111	111						
1524	SSR3996	D2	16	16	16						
	Satt050	A1	385	16	16						
	Satt432	C2	210	16	—						
	Sat_300	D2	16	5	—	16	16				
	Sat_338	D2	16	16	16	16	16	16	16	16	16
	Satt082	D2	16	16	16						
	Satt514	D2	16	16	16						
	Satt528	D2	16	52	16						
	Satt146	F	25	99	16	16	16	52	52	46	46
97	SSR0251	D1a	403	403	—						
	SSR0520	D1a	403	403	—						
	SSR2979	D1a	403	403	—						
	Satt129	D1a	54	403	403						
	Satt532	D1a	403	403	403	403	403				
	Satt242	K	386	403	403						

Note: Genomic and BAC end sequence from which microsatellites were derived was searched for sequence homology using BLAST against a preliminary 4× assembly of the whole genome sequence of soybean and the sequence contig identified was recorded. Only "best hits" are shown. Because Sat_ and Satt markers were screened against BAC pools they often identified multiple BACs, e.g., Sat_338. —, sequence not available.

[a] Marker WGS, whole genome sequence preliminary contig that was returned as a best hit when BLASTed against genomic DNA from which the microsatellite was derived.

[b] BES-F and BES-R, forward and reverse BES of BACs identified with microsatellites (Sat_ and Satt) or forward and reverse BES from which microsatellites were discovered (SSR).

[c] Because SSR markers were discovered in a BES the WGS contig for the marker is the same as the BES from which it was derived.

cause they also contained genetic "anchors" for which the marker map locations were contradictory, but at least one "anchor" consistent with the linkage group of the SSR mapped in this study. Examples of results from 3 FPC contigs are shown in Table 3. For example, for FPC contig 3955 the BES of the 3 BACs assayed all correspond to the same WGS contig, indicating that these BACs rightfully belong together in a contig. The sequences for 2 markers (SSR1766 and Satt268) match the same WGS contig, indicating that these markers are correctly associated with the FPC contig. Since both of these markers map to linkage group (LG) E, we can confidently say that FPC contig 3955 and WGS contig 111 are located on LG E. Satt542 is probably incorrectly associated with FPC contig 3955. In previous studies, markers occasionally identified multiple BACs in a contig (e.g., Sat_338 or Satt146). Our analyses sometimes yielded consistent results (Sat_338) and sometimes the BLAST results were ambiguous (Table 3).

These analyses confirmed the linkage group–FPC contig association for 30 (83%) of the 36 SSRs we examined (data not shown). The other 6 could not be confirmed by the BLAST results. The same analysis was able to identify 38 previously mapped microsatellites that were probably falsely associated with BACs within contigs. Results from this analysis also identified 4 FPC contigs that may contain improperly assigned BACs or, conversely, WGS assemblies that may require reassembly.

We also assayed BES and marker sequence from another 12 FPC contigs chosen at random that contained conflicting "anchors" but which did not contain any SSRs mapped in this study. Using the criteria listed we were able to identify the most probable correct anchor for 90% of the contigs (data not shown). A similar strategy is currently under way to assess all contig associations and the preliminary whole genome sequence assembly.

Discussion

More than 120 Mb of soybean genomic sequence derived from BAC end sequences were evaluated in this study. We found microsatellites with repeat lengths > 30 bp to be the most informative with our mapping parents and gel system. The frequency of polymorphisms remained relatively high between the interspecific parents and dramatically surged with repeat lengths greater than 100 bp. The frequency increased slowly in both intraspecific cross combinations until repeat lengths approached 60 bp. Length polymorphism frequency rose with very large repeat lengths in two populations, but the frequency dropped between the Minsoy and Noir I parents, suggesting perhaps common ancestry. We also identified repeat classes (e.g., AT, AG, AAT) that were strikingly more productive at identifying length polymorphisms than other classes (e.g., AAG or AGG). Some repeat classes, such as CCG, ACG, and AGC, were rarely observed. The SSRs identified, developed, and mapped in this study are available to researchers.

The frequency of length polymorphisms peaked at 15% in the Minsoy × Noir I population. This coincided with the frequency of polymorphisms detected in the same population by Shultz et al. (2007) but was considerably lower than the level reported by Song et al. (2004). Frequency of length polymorphism in the other two populations was also similar to that reported for other populations by Shultz et al. (2007) but was generally lower than the frequency detected in other studies using BARC_SSR markers (Njiti et al. 2002). This may be due to the methods of selection and development of the Beltsville, Maryland, BARC_SSRs, which imposed some selection for loci that were more polymorphic (Song et al. 2004).

A high-quality physical map is an important asset to a genomics research community, even a community with an impending whole genome sequence. An assessment of a strictly whole genome shotgun sequencing strategy for complex genomes predicted a number of potential problems, including a limited ability to resolve duplications, the concomitant loss of genes embedded within the duplicated segment, and an underestimation of the amount of euchromatin (She et al. 2004). A hybrid strategy involving resequencing of BACs mapped to the region of excessive divergence and read depth was recommended (She et al. 2004). The only way to prepare for this contingency is to have a well-developed physical map that is overlaid onto a sequence-based genetic map. Here we report the mapping of 265 SSRs derived from end sequences of BACs integrated into the soybean ‘Williams 82’ physical map.

There still remain numerous FPC contigs containing conflicting marker data. The ability of SSRs to detect single loci, even in a polyploid genome such as soybean, has made them essential in the creation of an integrated genetic linkage map (Cregan et al. 1999; Song et al. 2004). With this in mind, the assignment of microsatellite markers to BACs within a contig also containing microsatellite markers that map to alternative linkage groups remains a puzzle. During the development of the ‘Forrest’ physical map it was reported that many DNA markers anchored from 2 to 8 distinct contigs (Shultz et al. 2006). These authors indicated that SSRs were able to identify homeologous regions when used in BAC pools (Shultz et al. 2006). They concluded that each contig probably represented a homologous region of sequences derived from genome duplication events. Although we were unable to confirm this in the current study, this scenario is consistent with the relatively recent genome duplication that occurred in the soybean genome (Blanc and Wolfe 2004; Schlueter et al. 2004, 2006). It is also possible that microsatellites have been incorrectly assigned to specific BACs through false positives associated with multidimensional pooling strategies (Klein et al. 2000), cross-well contamination of libraries, etc.

Our results raise the number of genetically anchored contigs from 472 to 562. Because the SSRs used in this study were derived directly from BAC sequence, we can assume the marker–BAC associations are accurate to a high degree. The independent comparisons of BES and marker sequence to a preliminary WGS assembly also were shown to have the potential to increase the quality of contig assemblies. Here we used BLAST with BES and SSR-originating genomic sequence to look for matches with the preliminary 4× whole genome assembly of soybean. Some contigs resist analyses using this approach. This could be due to any number of reasons including a high degree of repetitive sequence, contig misassembly, etc. Although the WGS assembly is preliminary and is intended only for quality control evaluations, on the basis of the number of resolutions of incongruous BAC assignments, our results suggest that a similar genome-wide strategy may result in as much as a 90% improvement in physical map quality. It could also identify putative problem areas in FPC and WGS contigs. These efforts are currently under way. This version of the physical map is still undergoing independent review and quality control by members of the soybean genome community, and we expect the merging of additional contigs and the adjustment of contig order with more genetic marker information to be forthcoming.

Acknowledgements

This research was partially supported by grants from the United Soybean Board and the National Science Foundation (grant No. 0501877). We are grateful to DOE JGI for providing the preliminary 4× assembly of the soybean genome for these analyses. Names are necessary to report factually on the available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable.

References

Akkaya, M.S., Bhagwat, A.A., and Cregan, P.B. 1992. Length polymorphism of simple sequence repeat DNA in soybean. Genetics, **132**: 1131–1139. PMID:1459432.

Akkaya, M.S., Shoemaker, R.C., Specht, J.E., Bhagwat, A.A., and Cregan, P.B. 1995. Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci. **35**: 1439–1445.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Shang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-

BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402. doi:10.1093/nar/25.17.3389. PMID:9254694.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., et al. 2002. ARACHNE: A whole genome shotgun assembler. Genome Res. **12**: 177–189. doi:10.1101/gr.208902.

Blanc, G., and Wolfe, K.H. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell, **16**: 1667–1678. doi:10.1105/tpc.021345. PMID:15208399.

Burow, M.D., and Blake, T.K. 1998. Molecular tools for the study of complex traits. *In* Molecular dissection of complex traits. *Edited by* A.H. Paterson. CRC Press, Washington, D.C. pp. 13–29.

Cregan, P., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., and Kahler, A.L. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. **39**: 1464–1490.

Jackson, S., Rokhsar, D., Stacey, G., Shoemaker, R., Schmutz, J., and Grimwood, J. 2006. Toward a reference sequence of the soybean genome: a multiagency effort. Crop Sci. **46**: S-55–S-61.

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., et al. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. **13**: 91–96. doi:10.1101/gr.828403. PMID:12529310.

Klein, P.E., Klein, R.R., Cartinhour, S.W., Ulanch, P.E., Dong, J., Obert, J.A., et al. 2000. A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res. **10**: 789–807. doi:10.1101/gr.10.6.789. PMID:10854411.

Lewers, K.S., Crane, E.H., Bronson, C.R., Schupp, J.M., Keim, P., and Shoemaker, R.C. 1999. Detection of linked QTL for soybean brown stem rot resistance in 'BSR 101' as expressed in a growth chamber experiment. Mol. Breed. **5**: 33–42. doi:10.1023/A:1009634710039.

Lincoln, S.E., Daly, M.J., and Lander, S.L. 1993. Constructing genetic linkage maps with MAPMAKER/EXP version 3.0: a tutorial and reference manual. A Whitehead Institute of Biomedical Research Technical Report. 3rd ed. Whitehead Institute, Cambridge, Mass.

Luo, M.C., Thomas, C., You, F.M., Hsiao, J., Ouyang, S., Buell, C.R., et al. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics, **82**: 378–389. doi:10.1016/S0888-7543(03)00128-9. PMID:12906862.

Marek, L.F., and Shoemaker, R.C. 1997. BAC contig development by fingerprint analysis in soybean. Genome, **40**: 420–427. doi:10.1139/g97-056.

Morgante, M., Rafalski, A., Biddle, P., Tingey, S., and Olivieri, A.M. 1994. Genetic mapping and variability of seven soybean simple sequence repeat loci. Genome, **37**: 763–769. doi:10.1139/g94-109. PMID:8001811.

Nijiti, V., Meksem, K., Iqbal, M., Johnson, J., Kassem, M., Zobrist, K., et al. 2002. Common loci underlie field resistance to soybean sudden death syndrome in Forrest, Pyramid, Essex, and Douglas. Theor. Appl. Genet. **104**: 294–300. doi:10.1007/s001220100682.

Pampanwar, V., Engler, F., Hatfield, J., Blundy, S., Gupta, G., and Soderlund, C. 2005. FPC Web tools for rice, maize, and distribution. Plant Physiol. **138**: 116–126. doi:10.1104/pp.104.056291. PMID:15888684.

Peakall, R., Gilmore, S., Keys, W., Morgante, M., and Rafalski, A. 1998. Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol. Biol. Evol. **15**: 1275–1287. PMID:9787434.

Powell, W., Morgante, M., Doyle, J.J., McNicol, J.W., Tingey, S.V., and Rafalski, A.J. 1996. Genepool variation in genus *Glycine* subgenus *soja* revealed by polymorphic nuclear and chloroplast microsatellites. Genetics, **144**: 793–803. PMID:8889540.

Rozen, S., and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *In* Bioinformatics methods and protocols in the series methods in molecular biology. *Edited by* S. Krawetz and S. Misener. Humana Press, Totowa, N.J. pp. 365–386.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. Molecular cloning: a laboratory manual. 2nd ed. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.

Schlueter, J., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J., and Shoemaker, R. 2004. Mining EST databases to resolve evolutionary events in major crop species. Genome, **47**: 868–876. doi:10.1139/g04-047. PMID:15499401.

Schlueter, J., Scheffler, B., Schlueter, S., and Shoemaker, R. 2006. Sequence conservation of homeologous bacterial artificial chromosomes and transcription of homeologous genes in soybean (*Glycine max* L. Merr.). Genetics, **174**: 1017–1028. doi:10.1534/genetics.105.055020. PMID:16888343.

She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., et al. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature (London), **431**: 927–930. doi:10.1038/nature03062. PMID:15496912.

Shoemaker, R.C., and Olson, T.C. 1993. Molecular linkage map of the soybean (*Glycine max* L. Merr.). *In* Genetic maps: locus maps of complex genomes. *Edited by* S.J. O'Brien. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. pp. 6.131–6.138.

Shoemaker, R., Polzin, K., Labate, J., Specht, J., Brummer, E., Olson, T., et al. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics, **144**: 329–338. PMID:8878696.

Shultz, J.L., Kurunam, D., Shopinski, K., Iqbal, M.J., Kazi, S., Zobrist, K., et al. 2006. The soybean genome database (SoyGD): a browser for display of duplicated polyploid regions and sequence tagged sites on the integrated physical and genetic maps of *Glycine max*. Nucleic Acids Res. **34**: D758–D765. doi:10.1093/nar/gkj050. PMID:16381975.

Shultz, J.L., Kazi, S., Bashir, R., Afzal, J.A., and Lightfoot, D.A. 2007. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor. Appl. Genet. **114**: 1081–1090. doi:10.1007/s00122-007-0501-9. PMID:17287974.

Soderlund, C., Humphray, S., Dunham, A., and French, L. 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. **10**: 1772–1787. doi:10.1101/gr.GR-1375R. PMID:11076862.

Song, Q.J., Marek, L.F., Shoemaker, R.C., Lark, K.G., Concibido, V.C., Delannay, X., et al. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. **109**: 122–128. doi:10.1007/s00122-004-1602-3. PMID:14991109.

Stacey, G., Vodkin, L., Parrott, W.A., and Shoemaker, R.C. 2004. National Science Foundation-sponsored workshop report. Draft plan for soybean genomics. Plant Physiol. **135**: 59–70. doi:10.1104/pp.103.037903.

Wang, D., Shi, J., Carlson, S.R., Cregan, P.B., Ward, R.W., and Diers, B.W. 2003. A low-cost, high-throughput polyacrylamide gel electrophoresis system for genotyping with microsatellite DNA markers. Crop Sci. **43**: 1828–1832.

Warren, W.C., and The Soybean Mapping Consortium. 2006. A physical map of the "Williams 82" soybean (*Glycine max*) genome. Abstract W151. *In* Plant and Animal Genomes XIV Conference, San Diego, Calif., 14–18 January 2006.

Warren, R.L., Butterfield, Y.S., Morin, R.D., Siddiqui, A.S., Marra, M.A., and Jones, S.J.M. 2005. Management and visualization of whole genome assemblies using SAM. Biotechniques, **38**: 715–720. PMID:15945370.

Warren, R.L., Varabei, D., Platt, D., Huang, X., Messina, D., Yang, S.-P., et al. 2006. Physical map-assisted whole-genome shotgun sequence assemblies. Genome Res. **16**: 768–775. doi:10.1101/gr.5090606. PMID:16741162.

You, F.M., Luo, M.C., Gu, Y.Q., Lazo, G., Deal, K., Dvorak, J., and Anderson, O. 2007. GenoProfiler: batch processing of high-throughput capillary fingerprinting data. Bioinformatics, **23**: 240–242. doi:10.1093/bioinformatics/btl494. PMID:17018534.

Zou, J.J., Singh, R.J., Lee, J., Xu, S.J., Cregan, P.B., and Hymowitz, T. 2003. Assignment of molecular linkage groups to soybean chromosomes by primary trisomics. Theor. Appl. Genet. **107**: 745–750. doi:10.1007/s00122-003-1304-2. PMID:12783169.